جامعة السلطان قابوس
Sultan Qaboos University

College of Economics and Political Science

Information Systems Department

INFS6212 - Data Analytics for Business

Project Title:

Analyzing Community Interaction on Instagram using K-Means and Linear Regression.

Preparing By:

Mazin Alsalehi _ 147556

Moza Alkindi_ 147549

Bashayer Alhosni_ 147550

Submitted to:

Dr. Abdul Khalique Shaikh

Date:

10.05.2025

Table of Contents

1. Introduction 3
2. Brief about the Data Set 4
3. Methodology: 4
3.1 Steps to Apply Algorithms in RStudio 4
3.2 Segment Posts by Engagement Patterns using K-Means 4
3.3 Using Linear Regression to Predict Engagement: 7
4. Discussion and Results 9
4.1 Summary of Clustering Analysis on Instagram Community Interaction: 9
4.2 K-Means Clustering 10
4.3 Linear regression analysis 13
5. Key Findings from the Community Interaction Analysis 14
5.1 (K-Means Clustering) 15
5.2 (Linear Regression) 15
5.3 Strategic Value 16
6. Conclusion 17
7. References: 18

1. Introduction

In recent years, scientists have faced difficulty due to the growing diversity of social networks and their users. Finding similarities between people in these networks has become more popular as a result of the development of new interaction patterns and user grouping. Understanding the intricate and ever-changing structures of social networks is the goal of social network analysis, which includes this investigation as a crucial element.(Kanavos et al., 2023)

In digital communication, social media platforms are essential because they allow people and brands to interact with audiences, share content, and create online communities. The visually appealing format and variety of engagement capabilities, including likes, comments, shares, and saves, make Instagram stand out among these. For researchers and marketers, the abundance of user-generated content and interaction data offers a priceless chance to identify patterns in behavior and improve engagement tactics.

Through the analysis of post-level interaction data, this study seeks to investigate the connection between the behaviors of individual users and the larger community participation on Instagram. This is accomplished through the use of a dual-method analytical approach, which combines Linear Regression to forecast important factors influencing user reaction with K-Means Clustering to identify content groups based on engagement behavior.

Posts with comparable engagement characteristics are grouped using the unsupervised learning algorithm K-Means, which provides information on content categories that audiences respond to differently.(Song, 2016) Evidence-based content planning is made possible by the simultaneous modeling of how engagement factors like comments, shares, and saves affect the quantity of likes using Linear Regression, a fundamental supervised learning technique.

RStudio is used to perform the analysis, taking advantage of its powerful statistical and visualization features. This study offers both descriptive and predictive insights that enable data-driven decision-making in audience targeting, content optimization, and social media strategy by combining clustering and regression methodologies.

2. Brief about the Data Set

Instagram produces a significant amount of data each day. In order to find their postings, content creators on Instagram consider community involvement, engagement, and reach.The data set that will be used to analyze community interaction has 13 columns that record different interaction- and content-related measures, and it includes data from 119 Instagram posts. The columns include the origin of impressions (from Home, from Hashtags, from Explore, and from Other) and quantitative indicators like Impressions, Likes, Shares, Comments, Saves, Profile Visits, and Follows. Every post also has a caption and a collection of hashtags, which are both text data. There are no missing values in any of the data fields. The Caption and Hashtags fields are textual, although the majority of the columns are quantitative in nature. As indicated by the hashtags used, the posts' content frequently centers on subjects including finance, machine learning, data science, and Python programming. For instance, one post received 3,920 impressions, of which 1,028 came via hashtags and 619 from Explore. This resulted in 162 likes and 98 saves. A structured picture of Instagram users' interactions with content from various sources and themes is provided by this dataset.(Instagram Data, 2023)

3. Methodology:

3.1 Steps to Apply Algorithms in RStudio

K-Means Clustering and Linear Regression were used in the study's two-step data analytics methodology to examine Instagram community participation. RStudio's statistical and data visualization features were used to construct the tools and analysis. The information, which included 119 Instagram posts with metrics like

Like ...he pertinent
num ...rantee equal
influ

3.2

Step

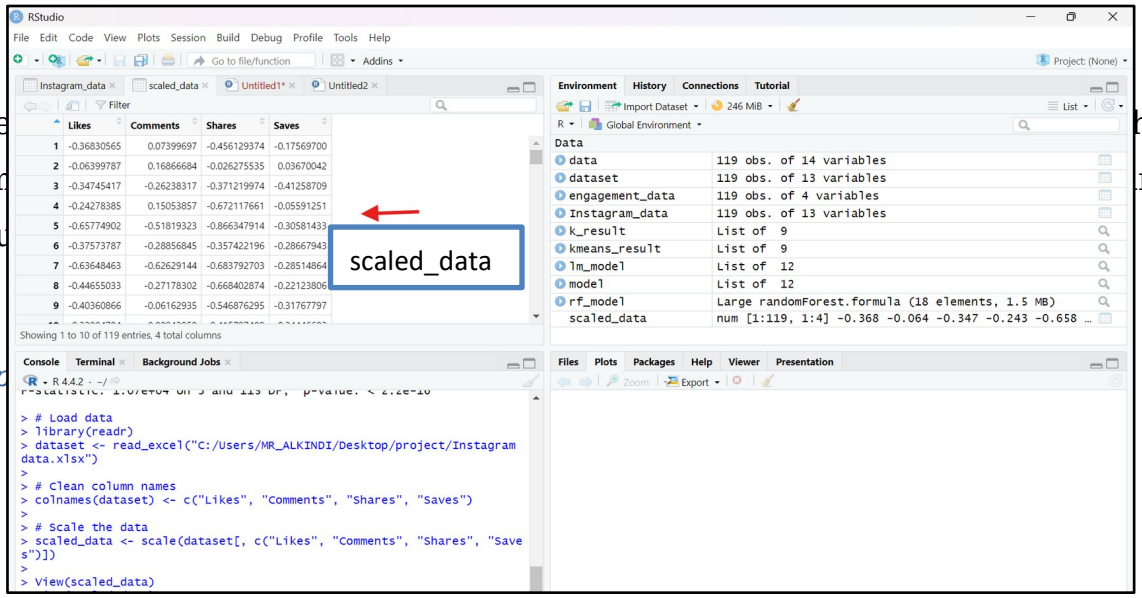


Figure 1: Step1_ Load and Prepare the Dataset

Step 2: Apply K-Means Clustering

To det ...l by plotting
the Wi ...elbow points
sugge ...timal k was
determ ... cluster was
then i

```
> # Load data
> library(readr)
> dataset <- read_excel("C:/Users/MR_ALKINDI/Desktop/project/Instagram data.xlsx")
> # Clean column names
> colnames(dataset) <- c("Likes", "Comments", "Shares", "Saves")
> # Scale the data (important for K-Means)
> scaled_data <- scale(dataset[, c("Likes", "Comments", "Shares", "Saves")])
> View(scaled_data)
```

```
# Load required libraries
library(factoextra) # for clustering visualization

# Elbow Method to choose optimal k (Figure 2)
fviz_nbclust(scaled_data, kmeans, method = "wss") +
  labs(title = "Elbow Method to Determine Optimal k")

# Based on elbow plot, choose k = 3
set.seed(123) # for reproducibility
kmeans_result <- kmeans(scaled_data, centers = 3, nstart = 25)

# Visualize clusters
fviz_cluster(kmeans_result, data = scaled_data) #(Figure 3)

fviz_cluster(k_result, data = scaled_data, geom = "point", ellipse.type =
"norm") +

  labs(title = "Post Segmentation by Engagement") # (Figure 4)
```

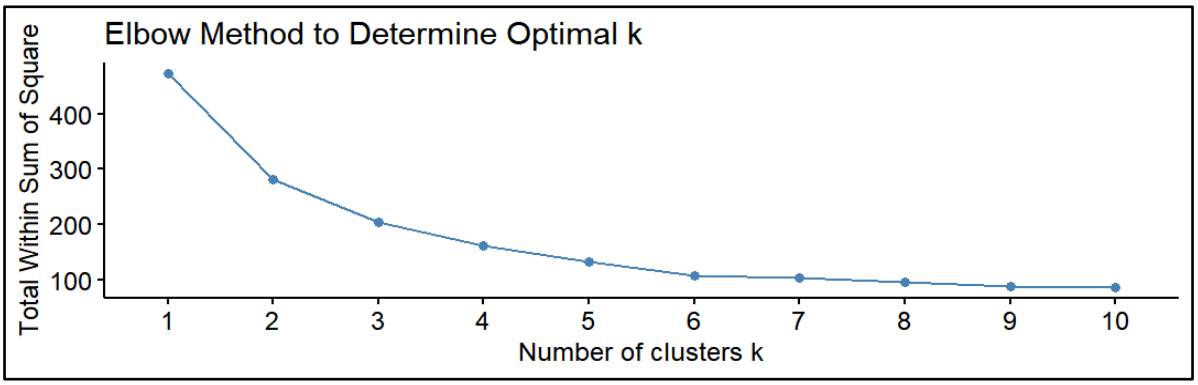


Figure 2: Elbow Method to choose optimal k

The Elbow Method (Figure 2) plotted the total within-cluster sum of squares (WSS) against different values of k to determine the ideal number of clusters (k). The "elbow," or the point at which the curve notably bends, is located at k = 3. Three clusters were chosen for the K-Means study because they offer the optimal balance

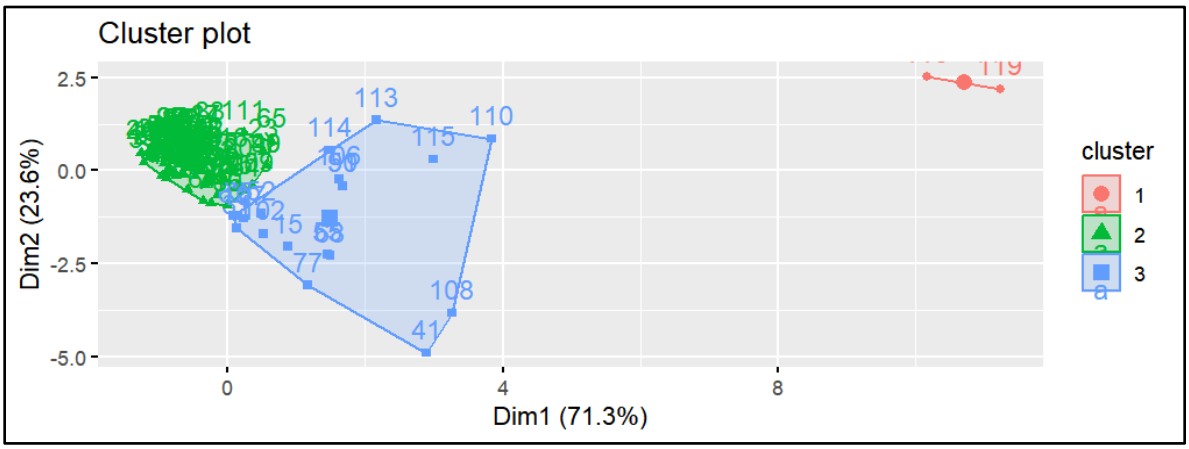
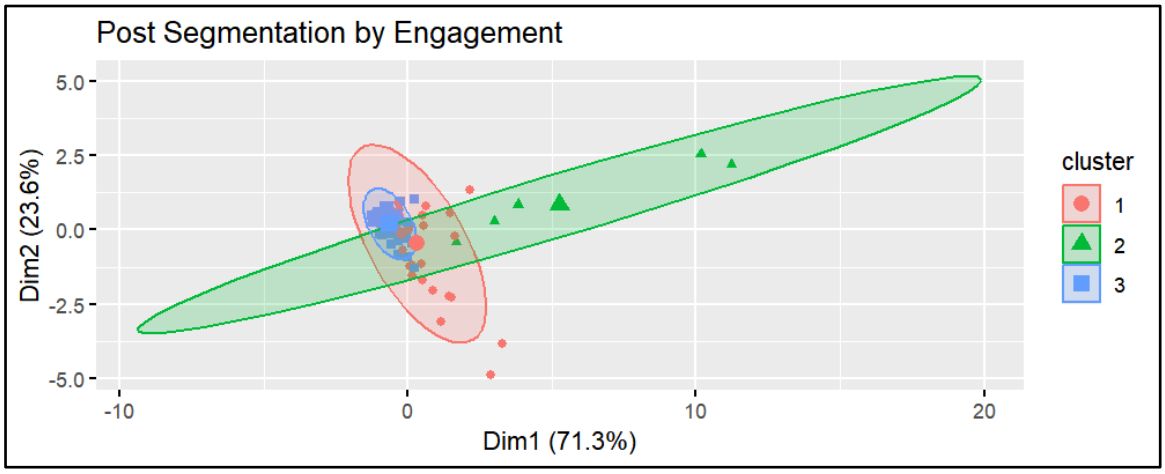


Figure3: Cluster Plot

between model simplicity and clustering effectiveness, according to this data.

This plot 

cluster

Figure4: Post Segmentation by Engagement

visualizes the output of the K-Means clustering. The plot is based on Principal Component Analysis (PCA), which reduces the dimensionality of the data (originally 4 metrics: Likes, Comments, Shares, Saves) into two principal components (Dim1 and Dim2) for clearer visualization. Here's a detailed explanation:

By applying K-Means clustering, it can turn raw engagement data into actionable content categories—enabling more focused, data-driven decisions in your Instagram strategy.

Based on engagement metrics, three different post groups were identified by cluster analysis. Clusters 1 and 3 demonstrate more similarity and denser grouping, but Cluster 2, which has a wider dispersion, illustrates postings with more variety or higher levels of interaction. A robust representation of the underlying data structure

was demonstrated by the first two dimensions, which accounted for almost 95% of the variance. A basis for focused content strategies catered to each engagement profile is provided by this segmentation.

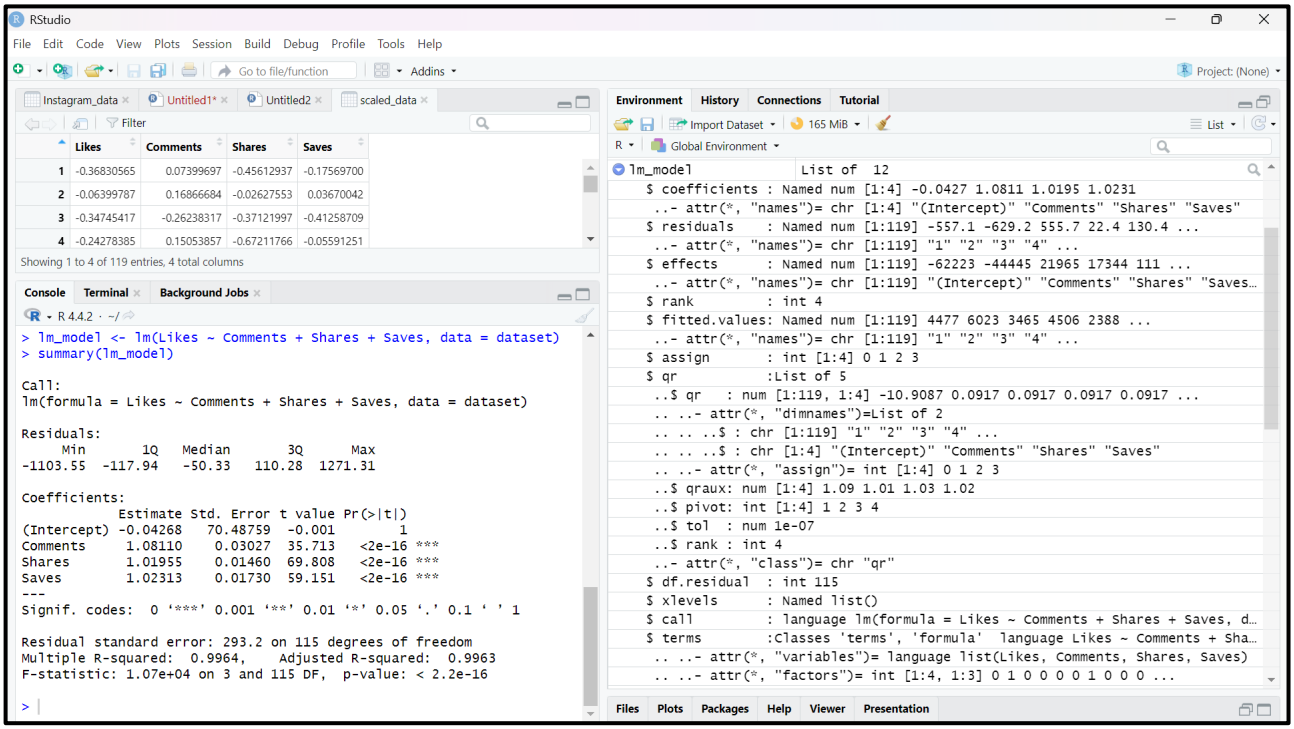
3.3 Using Linear Regression to Predict Engagement:

The purpose of linear regression is to understand how different features (e.g., Comments, Shares, Saves) relate linearly to a target metric like (Likes) and predict the number of Likes based on the other engagement types.

Step 3: Linear Regression to Predict Likes

```
# Fit linear regression model

lm_model <- lm(Likes ~ Comments + Shares + Saves, data = dataset)
summary(lm_model)
```


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Instagram_data
Untitled1*
Untitled2
scaled_data
Likes Comments Shares Saves
1 -0.36830565 0.07399697 -0.45612937 -0.17569700
2 -0.06399787 0.16866684 -0.02627553 0.03670042
3 -0.34745417 -0.26238317 -0.37121997 -0.41258709
4 -0.24278385 0.15053857 -0.67211766 -0.05591251
Showing 1 to 4 of 119 entries, 4 total columns
Console Terminal Background Jobs
R 4.4.2 · ~/
> lm_model <- lm(Likes ~ Comments + Shares + Saves, data = dataset)
> summary(lm_model)
Call:
lm(formula = Likes ~ Comments + Shares + Saves, data = dataset)
Residuals:
Min 1Q Median 3Q Max
-1103.55 -117.94 -50.33 110.28 1271.31
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.04268 70.48759 -0.001 1
Comments 1.08110 0.03027 35.713 <2e-16 ***
Shares 1.01955 0.01460 69.808 <2e-16 ***
Saves 1.02313 0.01730 59.151 <2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 293.2 on 115 degrees of freedom
Multiple R-squared: 0.9964, Adjusted R-squared: 0.9963
F-statistic: 1.07e+04 on 3 and 115 DF, p-value: < 2.2e-16
Environment History Connections Tutorial
Import Dataset 165 MiB
Global Environment
lm_model List of 12
$ coefficients : Named num [1:4] -0.0427 1.0811 1.0195 1.0231
$ residuals : Named num [1:119] -557.1 -629.2 555.7 22.4 130.4 ...
$ effects : Named num [1:119] -62223 -44445 21965 17344 111 ...
$ rank : int 4
$ fitted.values: Named num [1:119] 4477 6023 3465 4506 2388 ...
$ assign : int [1:4] 0 1 2 3
$ qr :List of 5
$ df.residual : int 115
$ xlevels : Named list()
$ call : language lm(formula = Likes ~ Comments + Shares + Saves, d...
$ terms :Classes 'terms', 'formula' language Likes ~ Comments + Sha...
Files Plots Packages Help Viewer Presentation

Figure 5: The result of linear regression

According to the results of the linear regression in (Figure 5), Comments, Shares, and Saves all contribute roughly one unit to the rise in Likes for every unit increase, making them statistically significant predictors of Likes ($p < 0.001$). Because of standardization, the intercept has no interpretive significance. With an R2 of 0.9964 and an adjusted R2 of 0.9963, the model shows excellent fit, meaning that the predictors account for more than 99% of the variance in Likes. Additionally, the model's overall strong statistical significance is confirmed by the F-statistic ($1.07e+04$, $p < 2.2e-16$).

St

```
# Generate predictions
predicted_likes <- predict(lm_model)
# Plot
plot(predicted_likes, dataset$Likes,
     main = "Predicted vs Actual Likes",
     xlab = "Predicted Likes", ylab = "Actual Likes",
     pch = 19, col = "steelblue")
abline(0, 1, col = "red")  # 45-degree reference line
```

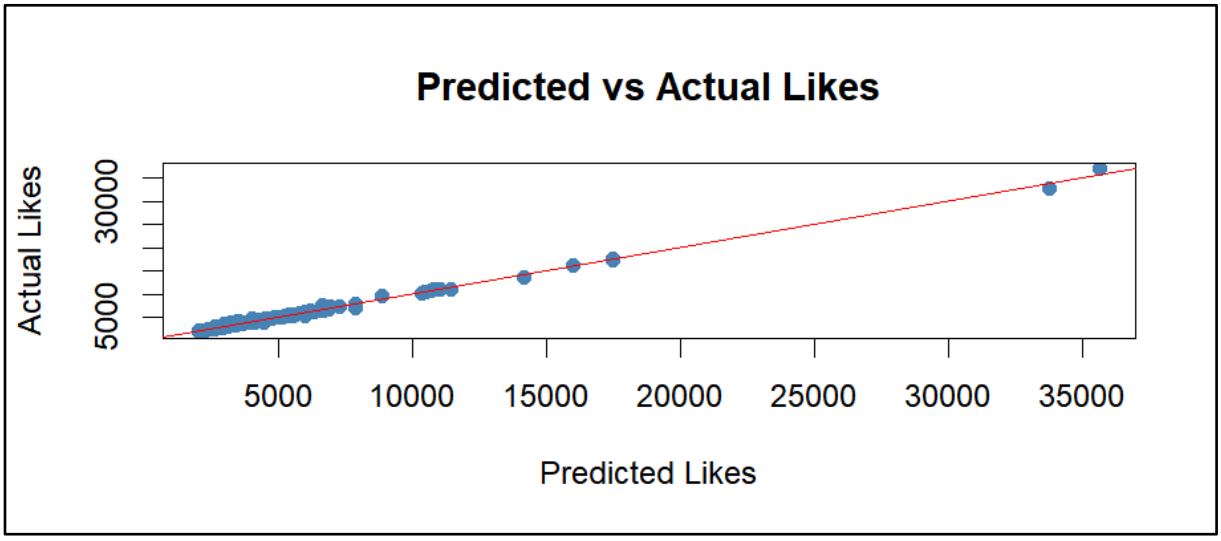


Figure 6: Predicted vs. Actual Likes Plot

The linear regression model (Likes ~ Comments + Shares + Saves) is diagnostically visualized by this "Predicted vs Actual Likes" scatter plot (Figure 6). It confirms the efficacy of the regression model. Predicted values closely match actual values, indicating that the model generalizes well to data and that Comments, Shares, and Saves are strong predictors of Likes.

As the chart shows:

- Each **blue dot** represents a data point (Instagram post).
- The x-axis shows the number of Likes predicted by your regression model.
- The y-axis shows the actual number of Likes from the dataset.
- The **red line** represents the ideal scenario where predicted Likes exactly match the actual Likes (a 45-degree line: y = x).

With the majority of dots closely fitting with the red reference line, the scatter plot comparing projected vs. actual Likes shows a strong linear relationship. This demonstrates the excellent explanatory power and precision of the linear regression model by showing that it predicts Likes with high accuracy and little residual error.Hence, the proposed algorithm is able to achieve good clustering performance. (Khan et al., 2021)

4. Discussion and Results

4.1 Summary of Clustering Analysis on Instagram Community Interaction:

This report is an analysis that uses data to find out the degree of the interactions in the community through Instagram through an approach called **K-Means Clustering**, which is an applied technique of data mining to partition your data set into clear and distinct levels or «clusters» based on similar features. And **linear regression analysis** was used to find out of all the types of interaction, the Comments, Shares, and Saves that correlated with the Likes and hence were major contributing factors to post popularity. The regression model examined the effect, both individually and collectively, of those three variables on the total Like count. The primary purpose of the study was to be able to recognize different forms of user interaction and get the necessary information to increase the effect of future content strategy.

4.2 K-Means Clustering

Data processing has brought in choosing and scaling the essential interaction variables. The set of data was made up of 119 Instagram posts with key interaction indicators. Prior to the clustering process, the numeric variables that were relevant were chosen and normalized. The number k of clusters had then been determined using the Elbow Method (Figure 2). This method allowed us to conclude that the optimal number of clusters was equal to three, as the inflection point was observed in the within-cluster sum of squares, which suggested that the data structure was intelligible with three clusters.

The three **result** clusters illustrated these engagement profiles in a clear way:

Cluster	Description	Interpretation	Strategic Insight
● Cluster 1 (Red – Top Right)	- Very few data points (e.g., post 119) - Extremely high values across engagement metrics - Considered outliers	Most viral or exceptional content — very rare but extremely high-performing	Analyze these posts to identify unique traits (e.g., topic, timing, format) to replicate success
▲ Cluster 2 (Green – Top Left)	- Majority of posts fall into this group - Moderate values across engagement metrics	Typical or average-performing content that moderately resonates with your audience	Use this cluster as your baseline and test optimization strategies to improve engagement
▨ Cluster 3 (Blue – Bottom Right)	- Distinct group with low engagement - Lower scores on both Dim1 and Dim2 axes compared to others	Underperforming content with low likes, shares, comments, and saves	Investigate these posts to identify weaknesses (e.g., poor timing, visuals) and consider revising

Table 1: Cluster Result

Strategic **Impact** of the Analysis:

- **Actionable Insights on Content Performance:** A good example is where the cluster analysis clearly separates the high-performing articles (Cluster 2) from the low-performing ones (Cluster 3), which offers benchmarks backed by data. The method becomes a reality check that gives you the possibility to compare the performance of your past posts based on the users' behavior at that time, as opposed to relying on subjective and unsupported beliefs.
- **Evidence-Based Strategy Optimization:** You can figure out which aspects lead to a more engaged audience (particularly Shares and Saves) audience by identifying the factors—like content type, visual style, posting time, and caption strategy—collecting

the corresponding values of these factors online and computing their correlations using open and free tools—and then you can create a list with the most efficient ones and the least effective ones that need to be improved or removed.

- **Data-Driven Content Calendar Planning:** After the key posts of high long-term benefit (e.g., Saves) and the ones of wide short-term appreciation (e.g., Shares) are known to you, you would then come up with a tailored content program that meshes with the desired campaign objectives—such as brand awareness, educational outreach, or sustained engagement.
- **Audience Engagement Profiling:** The segmented clusters illustrate the various points of engagement of the audience, thereby providing insights to how the different segments of your followers engage with content. The result is personalized content strategies, where posts are fitting to the targeted groups by selecting the right behavior of readers (e.g., savers, sharers, casual likers).

The segmentation uncovers important patterns that are used by the Instagram community when they interact with the content and presents several **benefits** in a practical sense:

- **Strategic Content Differentiation:** By recognizing post types that elicit high engagement (Clusters 1 and 2), content creators can formulate their approach to stress what appeals to their audience.
- **Improved Engagement Strategy:** Clusters having high performance provide the initial data on what the content formats, themes, or timing parameters of the community should be replicated or be created to get better interaction with the community in subsequent periods.
- **Content Performance Monitoring:** Regular employment of this cluster analysis technique offers a reliable and consistent way of tracking the shifts in engagement on a regular basis.
- **Insights into Community Behavior:** The findings from the study show that this tool enables a more personalized understanding of the target market's non-traditional touchpoints and behaviors; hence, this can be a valuable tool in making decisions for social media management.

Engagement with Instagram over a clustered approach has the advantage of segmenting shorter, more interactive content. The results could be useful in finding gaps in the content and initiating both audience and digital community participation in action (essay writing or reading, etc.).

- **K-Means Clustering limitations and recommendations**

No	Limitation	Recommendation
01	**Predefined Number of Clusters** In our analysis, we used the Elbow Method to find k. k is a number that refers to the number of clusters in a data set. The Elbow Method would tell us the value of k we actually need, viz. the least number of k that makes the high sum of squared errors (SSE) of the distance between data points from the centroid (within-cluster variance). When k is feasible and the best choice has been determined, the clusters will contain the most similar points.	Re-validate the choice of k in a number of ways over time like for instance using silhouette scores as an additional measure and especially getting insights from the field, thus leading to a more accurate determination of the number of groupings to choose the best fitting with the respective data set.
02	**Feature Scaling Sensitivity** We utilized clustering by various social media interactions such as Likes, Shares, Saves, and Comments. If these metrics had different ranges (were not standardized), the most significant variable (e.g., Likes) could have more impact than others on the	Always normalize the features (for example, by z-score transformation) before performing the clustering operation, keeping in mind that all the engagement factors are of equal importance and thus of one value to the cluster. Doing this will assign each post to the cluster closest to its centroid, which is

	result of the clusters.	the one of its categories.
03	**Rigid (Hard) Clustering** Perhaps a more appropriate model to be used in the future would be Gaussian Mixture Models as it allows some degree of overlapping the data by several clusters and can become the most flexible system of the available ones for detecting and holding simultaneous contributions of contents and cyclic appearances of the users.	Perhaps a more appropriate model to be used in the future would be Gaussian Mixture Models as it allows some degree of overlapping the data by several clusters and can become the most flexible system of the available ones for detecting and holding simultaneous contributions of contents and cyclic appearances of the users.
04	**Impact of Outliers** Viral or abnormally little-performing posts are at risk of changing the cluster centers and hence the model quality of segmentation.	Use outlier detection, for example boxplot or IQR method, and then delete or replace the extreme values to cluster only with the main features to make the clusters sturdy.

Table 2: K-Means Clustering limitations and recommendations

4.3 Linear regression analysis

A study of the Instagram engagement data was carried out to determine the interaction measurements that are highly related to post popularity measured in terms of the number of Likes. The respective predictor variables, i.e. Comments, Shares, and Saves, were used to figure out how much these factors contributed to the total Like count via the model.

The **result** indicated that Shares and Saves apparently were the main factors influencing Likes, thus they concluded that the act of saving and especially the act of sharing by users was really an essential part of the entire process. Comments added a positive effect, only not as appealing as the other two. The model's R-squared illustrated that a high percentage of

Likes' distribution could be statistically depicted by these three metrics, which gave a confirmation of the model's feasibility in the sphere of insight into engagement dynamics.

The **impact** of this analysis is that it helps us get to the bottom of the behavioral factors that guide people in their interaction. To be specific, the research demonstrates that content, the consumption of which by users results in deeper engagement—e.g., sharing with others or keeping for later—is sure to get more Likes. This information is very crucial for both community managers and content strategists since it furnishes them with a data-backed starting point for their posts which will encourage audience interaction and, as a consequence, the content will enjoy more reach and frequency

The **primary benefits** of this analysis involve better strategic planning for social media content, the possibility of predicting engagement performance, and the improved correlation between content creation and community behavior. In the last analysis, the results are the basis for better engagement strategies that can encourage a more interactive and faithful online community.

2. Linear Regression Limitations and Recommendations

No	Limitation	Recommendation
01	**Assumption of Linearity** In our model, we took Likes as a dependent variable and Shares, Saves, and Comments as independent variables. It is, for sure, a situation that the linear regression model assumes a linear relationship that can be a bad representation of real user engagement behavior.	To confirm the linearity of data, it can first plot the residuals and consider models such as Random Forest. The latter can potentially lead to a significant improvement in the model's performance. Thereafter, you can also plot the fitted values versus the residuals in order to detect an outlier in the model.
02	**Limited Prediction Accuracy— Linear regression** is the most	Utilize linear regression as the core or the initial model and then unleash

	explainable but cannot fully achieve the prediction goal in high-dimensional or non-linear data (e.g., Likes are also a function of other contributors, as well as a post schedule, hashtags, etc.).	advanced models (like Random Forest or XG Boost of machine learning models) to refine the predictions and incorporate interaction effects if any.

Table 3: Linear Regression Limitations and Recommendations

5. Key Findings from the Community Interaction Analysis

From the research that was carried out through K-Means Clustering and Linear Regression to study Instagram community interaction with the help of metrics such as Likes, Shares, Saves, and Comments:

5.1 (K-Means Clustering)

Cluster	Description
Cluster 1: Moderate Engagement	posts with rather low but numerous Saves and Shares and sort of Likes, maybe such content was informative or evergreen posts.
Cluster 2: High Engagement	posts with unprecedented Likes, Shares, and Saves, such posts are more likely to go viral or to belong to that kind of share.
Cluster 3: Low Engagement	posts with less activity over all the metrics, ones that would typically be the most suitable for the improvements or alternate solutions of content that were not followed initially.

Three different clusters of Instagram pics were spotted:

Table 4: K-Means Finding

Clustering has become a method of dividing content into significant categories, providing clarity to the team about the manner with which the various posts are liked by the audience. Visual methods (e.g., PCA plots, cluster centroids) ensured that the interpretation of the segmentation was feasible, identifying common patterns across post types.

5.2 (Linear Regression)

A linear regression model has been created to accurately conceptualize the Comments, Shares, and Saves and thus, decide the amount of Likes a post would obtain. The model states that both Shares and Saves are proven to be the most powerful and impactful drivers for generating Likes and engagement statistically.

With the model as a basis for the predictions, a few limitations were observed as a result of possible non-linear associations and multicollinearity among the factors. One of the primary results is that the findings are used to create content teams that generate engagement actions (e.g., getting more Saves) which are most likely to elicit higher Likes on the whole.

5.3 Strategic Value

The usage of both clustering and regression analysis served to deliver comprehensive as well as insightful feedback to the reader.

Approach	Value
Descriptive (K-Means)	What kind of posts are there, and what kind of engagement do users have with them?
Predictive (Regression)	Which of the factors can assist in the anticipation or reinforcement of the already achieved favorable engagement results?

Table 5: Strategic Value

By taking this coupled approach, we can ensure that the content is made in a more intelligent way, that the performance is upgraded, and that it is controlled in a more economical way (e.g., getting the most high-impact content made first would be the best option).

6. Conclusion

This project used a dataset of 119 posts and a dual-method data analytics approach—K-Means clustering and linear regression—to examine community interaction patterns on Instagram. The study effectively revealed important information about how people interact with content and the relationship between different engagement measures and post performance. Three significant groupings were formed from the postings using K-Means clustering:

- **High-performing content** with strong engagement across all metrics,
- **Informative content** with high saves and shares but moderate likes,
- **Low-engagement content**, indicating areas for improvement.

Based on audience behavior patterns, content makers can strategically adjust their approach with the help of these segments. Additionally, the regression model showed that the best predictors of Likes are Saves and Shares, underscoring the significance of content that entices people to share it right away or save it for later consideration. The R2 value of 0.9964 indicated that the model's predicted accuracy was outstanding.

Regression and clustering approaches work together to provide both descriptive and predictive information. This combined approach improves decision-making by offering a framework for audience targeting, performance tracking, and content improvement that is data driven.

Despite its effectiveness, the study has many drawbacks. Potential future research might examine non-linear models, include textual and visual content elements, and examine demographic data for more in-depth behavioral profiling.

In conclusion, this research shows how advanced analytics may convert unprocessed

engagement data into plans that can be implemented, allowing for more focused and effective social media efforts.

7. References:

1. Amir Motefaker. (2023). *Instagram Data*.

2. Kanavos, A., Karamitsos, I., & Mohasseb, A. (2023). Exploring Clustering Techniques for Analyzing User Engagement Patterns in Twitter Data. *Computers*, *12*(6). https://doi.org/10.3390/computers12060124

3. Khan, I., Luo, Z., Shaikh, A. K., & Hedjam, R. (2021). Ensemble clustering using extended fuzzy k-means for cancer data analysis. *Expert Systems with Applications*, *172*. https://doi.org/10.1016/j.eswa.2021.114622

4. Song, Y.-Tae. (2016). *2016 IEEE/ACIS 14th International Conference on Software Engineering Research, Management and Applications (SERA) : proceedings : June 8-10, 2016, Towson University, Balitmore, USA*. IEEE.